Montpellier, 13-15 June 2018

Book of Extended Abstracts

Comparing two models for disease mapping data not varying systematically in space

Helena Baptista[1,*], Jorge M. Mendes[1], Peter Congdon[2]

[1] *NOVA Information Management School, Universidade Nova de Lisboa, Lisboa, Portugal*
[2] *School of Geography and Life Sciences Institute, Queen Mary, University of London, UK*
[*] *NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal, mhbaptista@novaims.uml.pt*

Abstract. *Conditionally specified Gaussian Markov random field (GMRF) models with adjacency-based neighborhood weight matrix, commonly known as neighborhood-based GMRF models, have been the mainstream approach to spatial smoothing in Bayesian disease mapping. However, there are cases when there is no evidence of positive spatial correlation or the appropriate mix between local and global smoothing is not constant across the region being study. Two models have been proposed for those cases, a conditionally specified Gaussian random field (GRF) model using a similarity-based non-spatial weight matrix to facilitate non-spatial smoothing in Bayesian disease mapping, and a spatially adaptive conditional autoregressive prior model. The former model, named similarity-based GRF, is motivated for modeling disease mapping data in situations where the underlying small area relative risks and the associated determinant factors do not varying systematically in space, and the similarity is defined by similarity with respect to the associated disease determinant factors. In the presence of disease data with no evidence of positive spatial correlation, a simulation study showed a consistent gain in efficiency from the similarity-based GRF, compared with the adjacency-based GMRF with the determinant risk factors as covariate. The latter model considers a spatially adaptive extension of Leroux et al. [9] prior to reflect the fact that the appropriate mix between local and global smoothing may not be constant across the region being studied. Local smoothing will not be indicated when an area is disparate from its neighbours (e.g. in terms of social or environmental risk factors for the health outcome being considered). The prior for varying spatial correlation parameters may be based on a regression structure which includes possible observed sources of disparity between neighbours. We will compare the results of the two models.*

Keywords. *Bayesian modelling. Disease mapping. Small area.*

---

1 Introduction

Spatial disease mapping models are being extensively used to describe geographical patterns of mortality and morbidity rates. Information provided by these models is considered invaluable by health researchers and policy-makers as it allows, for example, to effectively allocate funds in high risk areas, and/or to plan for localised prevention/intervention programmes.

In cases of rare diseases and/or low populated areas, the classical estimators of the morbidity rates

show high variability, and spatial disease mapping models overcome that by borrowing strength from spatial *neighbours*. One rationale is that the spatial random effects used to implement such borrowing of strength are proxies for unobserved risk factors that vary smoothly in space. Models used in disease mapping are usually generalized linear mixed models (GLMM) formulated within a hierarchical Bayesian framework, and Poisson likelihood is often assumed for data in the form of counts of cases for each areal unit. Neighbourhood information is explicitly incorporated into the model by means of an appropriate prior specification. The seminal work of Besag et al. [3] provides a pair of area-specific random effects to model unstructured heterogeneity (extra-Poisson variation) and spatial similarity. The Besag-York-Mollié (BYM) model is an extension of the intrinsic conditional autocorrelation (CAR) model, a well known Gaussian Markov random field (GMRF) prior in disease mapping [3]. In the same field, Leroux et al. [9] proposed a conditional autoregressive prior incorporating a spatial correlation parameter, with its extreme values corresponding to pure spatial and pure unstructured residual variation. One important aspect of the CAR modelling is the definition of the so-called neighbourhood matrix, which characterizes the spatial structure of the data at hand, and is based on the concept of *neighbours*.

The debate on the definition of *neighbours* can be traced back to Besag [2]. Others have worked in defining *neighbours* in several different ways (Besag et al. [3], Best et al. [4], Earnest et al. [6], Congdon [5] and Lee and Mitchell [8]).

Most of the research in disease mapping is related with diseases resulting from environmental exposures, such as respiratory complications and cancer. Those extrinsic disease determinant factors are spatially smoothed, and using some kind of spatial proximity, either by adjacency or by distance, between areas in the definition of *neighbours* has therefore provided good results. In cases in which no spatial positive autocorrelation is displayed by the data, the neighbourhood matrix as it exists today may not be adequate. The similarity-based GRF approach, proposed by Baptista et al. [1], replaces the neighbourhood-based GMRF approach. The structure of the conditionals is maintained, but the smoothing and borrowing strength mechanisms are now based on the similarity of the areas, regardless of their relative location in space.

Another approach to the same aspect is proposed by Congdon [5], where is considered that uniform borrowing of strength based simply on proximity or contiguity may not be appropriate when there are discontinuities in the spatial pattern of health events or risk factors; for instance, a low mortality area surrounded by high mortality areas. Such discontinuity may often reflect spatial discontinuities in risk factors, whether observed or unobserved. An area showing such discontinuity may have a distorted smoothed rate when smoothing is towards the local mean.

In this submission we will present results of the implementation of the above two mentioned models. Section 2 will provide a brief overview of the similarity model, Section 3 will provide a brief overview of the adaptive model and Section 4 will provide a brief overview of the data used. This is still work in progress, so no results will be presented now.

2 A similarity-based Gaussian random field model

The GRF model proposed no longer retains the Markovian properties as those based on the neighbourhood weights. Instead of using spatial distance or spatial adjacency, a measure reflecting similarity between areas is introduced. Data used should come from: a) a disease determinant factor or a combination of factors, b) a source external to the survey that collected the disease data. The main objective of the

proposed model is the provision for borrowing strength between areas with similar disease determinant factors.

Firstly, regions exhibiting the *same or close* level of risk in a determinant factor will be regions with the *same or close* risk of the disease. Secondly, if disease data need to be *strengthened*, using disease determinant factor information collected by the same survey might inflate or not remediate possible *weaknesses* of the disease data. Therefore, an external source for the disease determinant factor is critical.

The rationale of our approach is the following: in cases of diseases with no environmental determinant factors, use of a positive spatial correlation based on physical distance or adjacency, in the GRF/GMRF model, may not be the best way to reflect similarity between areas. By using the GRF model reflecting *how similar* each area is to one another, in terms of a disease determinant factor that was collected by an external source, the disease risk distribution can be better assessed.

We use the BYM model (more detailed specifications can be found elsewhere [3]), with a neighbourhood matrix based on a matrix definition proposed by Best et al. [4], the new similarity matrix, with elements h_{ij} for each region i, with the following structure:

$$h_{ij} = \begin{cases} e^{-p_{ij}/b}, & \text{if } j \neq i \\ \frac{1}{n-1}\sum h_{(-i)}, & \text{otherwise,} \end{cases}$$

where p_{ij} is the absolute gap between region i and region j, $p_{ij} = |p_i - p_j|$, in terms of the disease determinant factor, and b is equal to a value that gives a relative weight of 1% ($h_{ij} = 0.01$) to an area i whose difference from an area j is the mean inter-region difference for the country. Elements h_{ii} need a specific definition, otherwise their value would be the one contributing the most to the prior, as $e^0 = 1$ and all other d_{ij} elements have values between 0 and 1. Therefore, p_{ii} values are equal to the average value of all elements except the ith area value.

More details can be found in Baptista et al. [1].

3 A Spatially Adaptive Conditional Autoregressive Prior

The similarity based GRF prior (Section 2) replaces spatial proximity as a basis for borrowing strength by similarity in one or more risk factors, and so takes explicit account of the actual spatial pattern of risk factors, allowing for the case when that pattern may be irregular (not spatially smooth). By contrast, the spatially adaptive approach retains the principle of spatial borrowing of strength, but modifies it to better represent discontinuities in the outcome and/or observed risk factors. The degree of spatial correlation is allowed to vary between sub-regions of the region under consideration, with one possible scheme linking varying spatial correlation to spatial similarity (or dissimilarity) in risk factors between an area and its surrounding locality.

We start with the Leroux et al. [9] model and here we propose spatial adaptivity based on area specific $\lambda \in [0,1]$, the uniform measure of spatial dependence. For areas i, distinctly low λ_i correspond to spatial disparate areas, unlike their neighbours in health risk and/or risk factors, so that there may be benefit in downweighting the principle of uniform pooling to the locality mean.

If predictors W_i measuring dissimilarity in observed risk factors are available, and so relevant to

whether pooling should be local or global, one can use a regression scheme $logit(\lambda_i) \sim N(W_i\gamma, 1/\tau_\lambda)$, where γ are regression parameters. For example, in Congdon (2008) the discrepancy measure is based on area socioeconomic deprivation z_i, with dissimilarity represented as $W_i = |z_i - \bar{Z}_i|$ with $\bar{Z}_i$ being average deprivation in the locality L_i around area i, namely $\bar{Z}_i = \sum_{j \in L_i} z_j/d_i$.

More details can be found in Congdon [5].

4 The data

The relative merits of the methodologies mentioned in sections 3 and 4 must be investigated by applying those models to data sets exhibiting different patterns of spatial association. Therefore, both models are assessed under the spatial association generated by data sets used either in Baptista et al. [1] (alcohol abuse disorder) and Congdon [5].

Other data sets are under investigation and may be used.

References

[1] Baptista, H., Mendes, J. M., MacNab, Y. C., Xavier, M., & de Almeida, J. M. C. (2016). A Gaussian random field model for similarity-based smoothing in Bayesian disease mapping. *Statistical Methods in Medical Research, 25(4)*, 1166-1184.

[2] Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society, 36*, 192–236.

[3] Besag, J., York, J., & Mollié, A. (1991). Bayesian Image Restoration, with Two Applications in Spatial Statistics. *Annals of the Institute of Statistical Mathematics, 43*, 1–20.

[4] Best, N.G, Arnold, R.A., Thomas, A., Waller, L.A. & Conlon, E.M. (1999). Bayesian Models for Spatially Correlated Disease and Exposure Data. *Bayesian Statistics 6* , 131–147.

[5] Congdon, P. (2008). A spatially adaptive conditional autoregressive prior for area health data. *Statistical Methodology, 5*, 552–563.

[6] Earnest, A., Morgan, G., Mengersen, K., Ryan, L. ,Summerhayes, R., & Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of Conditional Autoregressive (CAR) models. *International journal of health geographics, 6*, 54.

[7] Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and spatio-temporal epidemiology, 2*, 79–89.

[8] Lee, D., & Mitchell, R. (2013). Locally adaptive spatial smoothing using conditional auto-regressive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 62*, 593–608.

[9] Leroux, B. G., Lei, X., & Breslow, N. (2000). Estimation of Disease Rates in Small Areas: A New Mixed Model for Spatial Dependence. In M. E. Halloran and D. Berry (Eds.), *Statistical Models in Epidemiology, the Environment, and Clinical Trials* (Vol. 116, pp. 179-191). New York, NY: Springer New York.

[10] MacNab, Y. (2011). On Gaussian Markov random fields and Bayesian disease mapping. *Statistical methods in medical research, 20(1)*, 49–68

[11] Sun, D., Tsutakawa, R., & Speckman, P. L. (1999) Posterior distribution of hierarchical models using CAR distributions. *Biometrika, 86(2)*, 341-350.